



University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

9-1-2008

Statistical Inference in Dynamic Panel Data Models

Tze Leung Lai

Dylan S. Small
University of Pennsylvania

Jia Liu

Follow this and additional works at: http://repository.upenn.edu/statistics_papers



Part of the [Statistics and Probability Commons](#)

Recommended Citation

Lai, T., Small, D. S., & Liu, J. (2008). Statistical Inference in Dynamic Panel Data Models. *Journal of Statistical Planning and Inference*, 138 (9), 2763-2776. <http://dx.doi.org/10.1016/j.jspi.2008.03.011>

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/statistics_papers/578
For more information, please contact repository@pobox.upenn.edu.

Statistical Inference in Dynamic Panel Data Models

Abstract

Anderson and his collaborators have made seminal contributions to inference with instrumental variables and to dynamic panel data models. We review these contributions and the extensive economic and statistical literature that these contributions spawned. We describe our recent work in these two areas, presenting new approaches to (a) making valid inferences in the presence of weak instruments and (b) instrument and model selection for dynamic panel data models. Both approaches use empirical likelihood and resampling. For inference in the presence of weak instruments, our approach uses model averaging to achieve asymptotic efficiency with strong instruments but maintain valid inferences with weak instruments. For instrument and model selection, our approach aims at choosing valid instruments that are strong enough to be useful.

Keywords

instrumental variables, weak instruments, model selection, empirical likelihood, resampling

Disciplines

Statistics and Probability

Statistical inference in dynamic panel data models

Tze Leung Lai ^{*,1}

Department of Statistics, Stanford University, Stanford, CA, 94305-4065, USA.

Dylan S. Small

*Department of Statistics, The Wharton School, University of Pennsylvania,
Philadelphia, PA, 19104, USA.*

Jia Liu

*Department of Management Science and Engineering, Stanford University,
Stanford, CA, 94305-4026, USA.*

Abstract

T.W. Anderson and his collaborators have made seminal contributions to inference with instrumental variables and to dynamic panel data models. We review these contributions and the extensive economic and statistical literature that these contributions spawned. We describe our recent work in these two areas, presenting new approaches to (a) making valid inferences in the presence of weak instruments and (b) instrument and model selection for dynamic panel data models. Both approaches use empirical likelihood and resampling. For inference in the presence of weak instruments, our approach uses model averaging to achieve asymptotic efficiency with strong instruments but maintain valid inferences with weak instruments. For instrument and model selection, our approach aims at choosing valid instruments that are strong enough to be useful.

Key words: Instrumental variables; Weak instruments; Model selection; Empirical likelihood; Resampling.

1. Introduction and background

A panel (or longitudinal) data set follows the same sample of individuals over time, and thus provides multiple observations on each individual in the sample. A

* Corresponding author.

Email address: `lait@stanford.edu` (Tze Leung Lai).

¹ Supported by the National Science Foundation

panel data set combines the features of a cross-sectional and time-series data set and “allows economists and other social scientists to analyze, in depth, complex economic and related issues which could not have been treated with equal rigor using time series or cross-sectional data alone” (Hsiao, 1985). To illustrate the benefits of panel data, we describe the dynamic panel data model which was introduced by Balestra and Nerlove (1966) to study the demand for natural gas. In the residential and commercial markets for heating fuel, a short-run change in the relative price of natural gas does not induce many consumers to immediately revise their consumption choices because of the high transfer costs involved in the shift to a different fuel. The consumption of gas is to a large extent governed by the stock of gas appliances and the price of gas which vary over time. Balestra and Nerlove proposed the model

$$y_{it} = \beta y_{i,t-1} + \delta' z_i + \gamma' x_{it} + v_{it}, \quad (1)$$

where $i = 1, \dots, n$ are individual units (U.S. states in their study), $t = 1, \dots, T$ are time periods, y_{it} is the amount of natural gas consumed by unit i at time t , z_i is a vector of time-invariant covariates, x_{it} is a vector of time-varying covariates (e.g., the price of gas), and v_{it} is a structural disturbance that is decomposed into a time-invariant part α_i , called the individual effect, and a time-varying part u_{it} . The parameters δ , γ and β have a causal interpretation – they represent the expected effect on y_{it} of changing the covariate z_i or x_{it} , or $y_{i,t-1}$ and leaving all other included and omitted variables constant. The disturbance v_{it} includes the effect on y_{it} of all variables not included in the model as well as random factors. Consequently, v_{it} may be correlated with x_{it} , z_i and $y_{i,t-1}$. Note that the model distinguishes between the short-run effect (γ) of a change in x_{it} and the long-run effect, $\gamma/(1 - \beta)$ if $|\beta| < 1$, of a change in x_{it} .

To focus on the basic statistical issues concerning (1), Anderson and Hsiao (1981) considered the simpler model without covariates:

$$y_{it} = \beta y_{i,t-1} + \alpha_i + u_{it}, \quad (2)$$

with the following basic assumptions:

- (A1) The random vectors $(y_{i0}, \dots, y_{iT}; u_{i0}, \dots, u_{iT}; \alpha_i)$ are i.i.d.
- (A2) $E(\alpha_i) = 0$, $E(u_{it}) = 0$, $E(\alpha_i u_{it}) = 0$ for $t = 1, \dots, T$.
- (A3) $E(u_{is} u_{it}) = 0$ for $s \neq t$.
- (A4) $E(u_{it} y_{i0}) = 0$ for $t = 1, \dots, T$.

Even if we make further distributional assumptions about the disturbances u_{it} and α_i , model (2) and assumptions (A1)-(A4) do not fully describe the stochastic process (y_{i0}, \dots, y_{iT}) because they do not describe how the y_{i0} are generated at time 0. Different assumptions about how the y_{i0} are generated alter the likelihood function and the interpretation of how y_{it} is generated. In conventional time series analysis, assumptions about how the initial conditions are generated are usually innocuous because T is assumed to be large. However, in economic panel data sets, T is usually small and N is large (Chamberlain, 1984). Anderson and Hsiao (1981) derived the

maximum likelihood estimator of the model (2) under various assumptions about the initial conditions and the assumption of normally distributed α_i and u_{it} . In the asymptotic regime in which T is fixed and $N \rightarrow \infty$, they showed that the MLE's based on different assumptions about the initial conditions are not asymptotically equivalent and may be inconsistent if the initial conditions are misspecified. However, as they noted, "we usually have little information to rely on in making a correct choice of the initial conditions." They therefore proposed to difference the data, thereby eliminating the α_i 's and then to use the differenced data to estimate β . In particular, taking first differences yields

$$y_{it} - y_{i,t-1} = \beta(y_{i,t-1} - y_{i,t-2}) + u_{it} - u_{i,t-1}. \quad (3)$$

If one estimates β in (3) by least squares, the estimate is inconsistent for T fixed, $n \rightarrow \infty$, because $\text{Cov}(y_{i,t-1} - y_{i,t-2}, u_{it} - u_{i,t-1}) \neq 0$. As noted by Anderson and Hsiao (1981), multiplying both sides of (3) by $y_{i,t-2}$ gives

$$y_{i,t-2}(y_{it} - y_{i,t-1}) = \beta y_{i,t-2}(y_{i,t-1} - y_{i,t-2}) + y_{i,t-2}(u_{it} - u_{i,t-1}). \quad (4)$$

Because $E(u_{it}y_{i0}) = 0$ and the u_{it} 's are serially uncorrelated and uncorrelated with α_i under the assumptions (A2)-(A4), $E[y_{i,t-2}(u_{it} - u_{i,t-1})] = 0$. Hence, one can use (4) to estimate β via the method of moments:

$$\hat{\beta} = \frac{\sum_{i=1}^N \sum_{t=2}^T (y_{it} - y_{i,t-1})y_{i,t-2}}{\sum_{i=1}^N \sum_{t=2}^T (y_{i,t-1} - y_{i,t-2})y_{i,t-2}}. \quad (5)$$

The estimate (5) is often called an *instrumental variable* estimate with $y_{i,t-2}$ as the instrument. By the law of large numbers, this estimate is consistent for T fixed and $n \rightarrow \infty$ if $\beta \neq 0$. Anderson and Hsiao (1981) noted that one could alternatively use $y_{i,t-2} - y_{i,t-3}$ as an instrument to estimate β by

$$\hat{\beta} = \frac{\sum_{i=1}^N \sum_{t=3}^T (y_{it} - y_{i,t-1})(y_{i,t-2} - y_{i,t-3})}{\sum_{i=1}^N \sum_{t=3}^T (y_{i,t-1} - y_{i,t-2})(y_{i,t-2} - y_{i,t-3})}. \quad (6)$$

In fact, the basic assumptions (A1)-(A4) about (2) imply $T(T-1)/2 + (T-2)$ moment restrictions:

$$E[y_{is}(u_{it} - u_{i,t-1})] = 0 \text{ for } t = 2, \dots, T; \quad s = 0, \dots, t-2, \quad (7)$$

$$E[(\alpha_i + u_{iT})(u_{it} - u_{i,t-1})] = 0 \text{ for } t = 2, \dots, T-1. \quad (8)$$

Each restriction provides an estimating equation that can be used to construct a consistent estimate of β by the method of moments.

Anderson and Hsiao (1982) have extended this approach to the more general model (1) with covariates for which there are more possible assumptions about the relationships between a covariate x_{it} (or z_i) and the disturbance terms. In particular, they consider the following important examples:

(a) x_{it} is predetermined with respect to u_{it} so that

$$E(u_{it}x_{is}) = 0, \quad s = 1, \dots, t, \quad t = 1, \dots, T$$

(b) x_{it} is strictly exogenous with respect to u_{it} so that

$$E(u_{it}x_{is}) = 0, \quad s = 1, \dots, T, \quad t = 1, \dots, T$$

(c) x_{it} is uncorrelated with the individual effect α_i , i.e.,

$$E(x_{it}\alpha_i) = 0, \quad t = 1, \dots, T.$$

Each of these assumptions provides a series of estimating equations from which (β, δ, γ) can be identified and estimated. Bhargava (1991b) provides sufficient conditions for the identification of dynamic panel data models containing endogenous variables. Identification can be achieved under quite weak conditions even in the presence of a general pattern of correlation between the disturbances v_{it} and the time-varying variable and therefore it is not necessary to make the restrictive assumption that the time-varying parts of v_{it} are serially uncorrelated. The dynamic panel data model (1) has been widely used in empirical economic studies. Examples include Bhargava (1991a), who studied the income elasticity of the demand for food and nutrients in rural South India using a panel of households, Baltagi and Levin (1986), who studied the effects of taxation and advertising on cigarette demand using a panel of U.S. states, Lee, Longmire, Mátyás and Harris (1998), who studied convergence of countries' economic outputs and Pesaran, Haque and Sharma (2000), who studied the determinants of economic growth using a panel of countries.

The instrumental variable (IV) method proposed by Anderson and Hsiao (1981) and subsequent refinements that combine multiple instruments efficiently have been widely used to estimate dynamic panel data models. The IV method is also used for making inferences in many other complex economic models. Although consistency and asymptotic normality of the IV method has been established under mild regularity conditions, there is considerable evidence that this asymptotic theory provides a poor approximation to the sampling distributions of IV estimators and associated test statistics in many designs and sample sizes of empirical relevance in economics. The July, 1996 issue of the *Journal of Business and Economic Statistics* is devoted to this topic. Examples of the large discrepancy between asymptotic theory and finite sample performance have been well documented for both linear instrumental variable regression models (Anderson and Sawa, 1973, 1979; Nelson and Startz, 1990a,b; Bound, Jaeger and Baker, 1995; Staiger and Stock, 1997) and dynamic panel data models (Blundell and Bond, 1998; Alonso-Borrego and Arellano, 1999). The sampling distribution of IV estimators can be skewed and have heavy tails. In addition, tests of the parameter values and overidentifying restrictions can exhibit substantial size distortions. Many of these problems can be traced to the presence of "weak" instruments and/or a large number of instruments. A rough definition of a weak instrument is that the associated moment restriction provides little information relative to the sample size. A particularly striking example of the problems that weak

instruments can cause was highlighted by Bound, Jaeger and Baker (1995). They re-examined Angrist and Krueger’s (1991) study of the causal effect of schooling on earnings which used over 329,000 observations. Bound, Jaeger and Baker showed that the first-order asymptotic theory for the asymptotically efficient IV estimator was unreliable even for this large sample size because the instrument was extremely weak. Other economic studies in which weak instruments have been found to cause problems include studies of intertemporal labor supply (Lee, 2001), returns to scale in industry (Burnside, 1996) and asset pricing (Stock and Wright, 2000). Weak instruments are common in economic studies using instrumental variables to make causal inferences from observational studies because unless there is an extremely fortunate natural experiment, any variable that satisfies the moment restriction is unlikely to have a strong correlation with the endogenous variables.

Inference in the presence of weak instruments, therefore, has been an important area of research in econometrics. Seminal contributions by Anderson and Rubin (1949, 1950) and Anderson (1950) show that valid parametric inference can still be performed in the presence of weak instruments by making use of limited-information maximum likelihood (LIML). As pointed out by Andrews, Moreira and Stock (2006), “except for the important early contribution by Anderson and Rubin,” most of this literature is “recent” and appeared in the past decade. Section 2 reviews these recent developments that are parametric in scope, following the seminal work of Anderson and Rubin. It also describes a new nonparametric approach that uses empirical likelihood to replace parametric likelihood, and resampling to circumvent the inadequacy of asymptotic approximations, either normal or non-Gaussian, in the presence of weak instruments.

As pointed out in (7) and (8) and their succeeding paragraph, one often comes up with a large number of instruments in using the IV method, leading to an overdetermined system with more estimating equations than parameters. In addition, when there are exogenous variables, one also encounters the issue of variable selection. Section 3 reviews previous approaches to model and moment selection related to the IV method and describes a new approach that is targeted towards the actual use of the selected model. An example from dynamic panel data models is used to illustrate the approach.

Amemiya (1990) comments that “there have been many applications and generalizations of the Anderson-Hsiao formulations both in the econometric and statistical literature, indicating their far-reaching impact on the profession.” The earlier work of Anderson and Rubin provides the foundation for valid inference when the Anderson-Hsiao approach ends up with weak instruments. As indicated in Sections 2 and 3, these seminal ideas have spawned active areas of research in econometrics and statistics.

2. Valid inferences in the presence of weak instruments

2.1 Generalized method of moments and empirical likelihood

In Section 1, we have noted that for the dynamic panel data model, there are several IV estimators, e.g., (5) and (6), each based on a different moment restriction

(estimating equation). This section describes two basic approaches, the generalized method of moments and empirical likelihood, to combining the information in the available moment restrictions to estimate β in the following general class of models specified by moment restrictions. Let X, X_1, \dots, X_n be i.i.d. random vectors having unknown distribution function F . Suppose that we are interested in F through the function $\beta = \beta(F)$ where $\beta \in \mathcal{B}$ is p -dimensional and (F_0, β_0) denotes the true value of (F, β) . The information about β and F is available in the form of $r \geq p$ functionally independent moment restrictions:

$$E_F[g(X, \beta_0)] = 0, \quad (9)$$

$$g(X, \beta) = (g_1(X, \beta), \dots, g_r(X, \beta))'.$$

It is also assumed that $\beta_0 = \beta(F_0)$ is the unique value of β satisfying (9), i.e.,

$$E_{F_0}[g(X, \beta)] \neq 0 \quad \text{if } \beta \neq \beta_0. \quad (10)$$

This condition is necessary for β to be identifiable. Let \mathcal{F} denote the family of distributions satisfying (9) for some β . For the dynamic panel data model (2) without covariates, $X_i = (y_{i0}, \dots, y_{iT})$ and the moment restrictions are given by (7)-(8). Semiparametric models of the form (9) are used in many economic studies because economic theory often provides information about certain aspects of the data generating process but does not specify an exact distributional form for the data generating process (Harris and Mátyás, 1999). If $r = p$, the sample version of (9) gives a *just-identified* system of estimating equations whose solution corresponds to the estimator of β obtained by the method of moments. If $r > p$, it gives an *over-identified* system of estimating equations that need to be combined in some way to estimate β .

Hansen (1982) coined the term *generalized method of moments* (GMM) for the asymptotically optimal way to combine estimating equations; Chiang (1956) and Ferguson (1958) showed related results. Let $G(\beta) = E_{F_0}[g(X, \beta)]$, whose sample version is $\hat{G}(\beta) = \frac{1}{n} \sum_{i=1}^n g(X_i, \beta)$. Hansen (1982) introduces a positive definite weight matrix W_n and minimizes the quadratic form $\hat{G}'(\beta)W_n\hat{G}(\beta)$. The GMM estimator is defined by

$$\hat{\beta}_n = \arg \min_{\beta} \hat{G}'(\beta)W_n\hat{G}(\beta). \quad (11)$$

Equivalently, the GMM estimator solves the estimating equation

$$\frac{\partial \hat{G}(\beta)'}{\partial \beta} W_n \hat{G}(\beta) = 0. \quad (12)$$

Under certain regularity conditions, the GMM estimator $\hat{\beta}_n$ is consistent and asymptotically normal, and an asymptotically optimal choice of W is $V = [\text{Cov}(g(X, \beta_0))]^{-1}$. Since V is unknown, Hansen (1982) proposes to start with a positive definite matrix

W , e.g., the identity matrix, to obtain first a consistent estimator $\tilde{\beta}_n$ of β_0 . Then $\text{Cov}(g(X, \beta_0))$ can be estimated by the sample covariance matrix with β_0 replaced by $\tilde{\beta}_n$. The GMM estimate $\hat{\beta}_n$ is then obtained by (11) with $W_n = [\hat{\text{Cov}}(g(X, \tilde{\beta}_n))]^{-1}$. This two-step GMM estimator is consistent, as shown by Hansen (1982). It is asymptotically equivalent to the maximum empirical likelihood estimator described below.

For parametric problems, basing inference on the likelihood is an effective and flexible general principle that has asymptotically optimal properties. Owen (1988) has extended the scope of likelihood inference to nonparametric functionals by using empirical likelihood. For model (9), the nonparametric likelihood of $(F, T(F))$ is

$$L(F, T(F)|X_1, \dots, X_n) = \prod_{i=1}^n F(\{X_i\}),$$

where $\beta = T(F)$ and $F(\{X_i\})$ denotes the probability of obtaining the value X_i in a sample from F . The empirical likelihood of β is defined as the profile likelihood for β , i.e.,

$$L_E(\beta) = \max_{F, T(F)=\beta} L(F, T(F)).$$

Note that $L(F|X_1, \dots, X_n) = 0$ if the support of F is larger than that of the empirical distribution. This means that \hat{F}_β , the maximum likelihood estimate of F restricted to $\{F : T(F) = \beta\}$, must have $\{X_1, \dots, X_n\}$ as its support. Thus, \hat{F}_β is a multinomial distribution on X_1, \dots, X_n with probabilities p_1, \dots, p_n and we can write the empirical likelihood of β as

$$L_E(\beta) = \sup \left\{ \prod_{i=1}^n p_i \mid p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g(X_i, \beta) = 0 \right\}. \quad (13)$$

Let $l_E(\beta) = \log L_E(\beta)$. Owen (1988) and Qin and Lawless (1994) have shown that for model (9), $2\{l_E(\hat{\beta}_{MELE}) - l_E(\beta_0)\} \Rightarrow \chi_p^2$, where $\hat{\beta}_{MELE}$ is the maximum empirical likelihood estimator. This Wilks-type theorem provides asymptotically valid $100\alpha\%$ empirical likelihood confidence regions, $\{\beta : 2l_E(\hat{\beta}_{MELE}) - 2l_E(\beta) \leq \chi_{p, 1-\alpha}^2\}$, where $\chi_{p, 1-\alpha}^2$ is the $(1 - \alpha)$ -quantile of the χ_p^2 distribution. Moreover, the MELE has the same first-order asymptotic properties as the GMM estimator of β in model (9):

$$\sqrt{n}(\hat{\beta}_{MELE} - \beta_0) \Rightarrow N(0, (D'V^{-1}D)^{-1}), \quad (14)$$

where $D = E \left\{ \frac{\partial g(X_i, \beta)}{\partial \beta} \right\}$. Note that (14) corresponds to the asymptotic theory of Wald's statistic in parametric inference.

2.2 Weak instruments, LIML and conditional likelihood ratio tests

A moment restriction or instrumental variable is said to be *weak* if it provides little information, relative to the sample size, on the unknown parameter vector β . Because weak instruments result in poor asymptotic approximations for IV- or

GMM-based inference but occur commonly in economic studies, as noted in Section 1, there has been considerable interest in how to make valid inferences in a setting of weak instruments. The seminal work of Anderson and Rubin (1949, 1950) in this area is concerned with inference on β in the linear IV regression model

$$y_i = \beta' w_i + \gamma' x_i + u_i, \quad (15)$$

$$E(u_i x_i) = 0, E(u_i z_i) = 0, E(u_i) = 0, \text{Var}(u_i) = \sigma_u^2. \quad (16)$$

in which the endogenous variables w_i satisfy the linear regression relationship

$$w_i = \Pi_1 x_i + \Pi_2 z_i + \xi_i, \quad E(\xi_i x_i) = 0, E(\xi_i z_i) = 0 \quad (17)$$

i.e., $E(w_i | x_i, z_i)$ is linear in x_i and z_i . In order for β to be identifiable, we assume that $\dim(z_i) \geq \dim(\beta)$; if $\dim(z_i) > \dim(\beta)$, β is said to be overidentified. Let $\dim(z_i) = m_2$ and $\dim(\beta) = m_1$; therefore Π_2 is $m_1 \times m_2$ with $m_1 \leq m_2$. The LIML (*limited-information maximum likelihood*) approach of Anderson and Rubin (1949, 1950) is based on the simple idea that if β is specified, the linear structural model can be reduced to a simple linear regression equation. In particular, if we consider the hypothesis $H_0 : \beta = \beta_0$ in model (15)-(17), we can write

$$y_i - \beta_0' w_i = \theta_1' x_i + \theta_2' z_i + \epsilon_i, \quad (18)$$

where $\theta_1 = \gamma + \Pi_1'(\beta - \beta_0)$, $\theta_2 = \Pi_2'(\beta - \beta_0)$, and $\epsilon_i = u_i + (\beta - \beta_0)' \xi_i$. Under H_0 , (18) satisfies all the conditions of the linear regression model with $\theta_2 = 0$. The F -statistic for testing $H_0' : \theta_2 = 0$ in the regression model (18) has an asymptotic χ_s^2/s null distribution for any error distribution of (u, ξ) as $n \rightarrow \infty$, where $s = \dim(z)$. Since this asymptotic null distribution does not depend on Π_2 , the Anderson-Rubin test retains asymptotically valid levels of significance under weak instruments.

The single structural equation (18) is also referred to as the *Limited Information Simultaneous Equations Model* (LISEM) as it can be thought of as being selected from a simultaneous system (15)-(17); see Staiger and Stock (1997). Assuming normality, Anderson and Rubin (1949) have derived the maximum likelihood estimator, called LIML, in this model. Under the additional assumption of known error covariance matrix, Moreira (2003, pp. 1030-1031) constructs sufficient statistics for the model and uses invariance with respect to a group of linear transformations to further reduce dimensionality, yielding a bivariate invariant sufficient statistic which has the Anderson-Rubin statistic as one of the components. In the just-identified case, the invariant sufficient statistic is the Anderson-Rubin statistic, and the Anderson-Rubin test is optimal (i.e., uniformly most powerful) among invariant similar tests. However, in the over-identified case, the Anderson-Rubin test may have poor power because it ignores the other component, denoted by T , of the bivariate invariant sufficient statistic as pointed out by Moreira (2003). In the over-identified case, Moreira proposes a *conditional likelihood ratio* (CLR) test, which is a similar test whose critical value is determined by the conditional null distribution of the likelihood ratio

statistic given T . The conditional null distribution does not depend on the unknown parameters.

As pointed out by Moreira (2003, *p.* 1032), the asymptotic null distribution of the likelihood ratio statistic LR under $\beta = \beta_0$ depends on an unknown parameter τ that ranges between 0 and ∞ , making LR non-pivotal. On the other hand, Wang and Zivot (1998) note that under the null hypothesis, LR is stochastically smaller than a pivotal statistic, which they use to obtain a valid but conservative test. Kleibergen (2002) proposes a Lagrange Multiplier (LM) test which is asymptotically pivotal. Moreira (2003) and Andrews, Moreira and Stock (2006, 2007) have carried out simulation studies which show the CLR test to outperform the other tests. They have also extended the CLR test to non-normal errors and unknown covariance matrices by treating the errors as if they were normal and replacing the covariances by their sample estimates, using both weak and strong instrument asymptotics to give an asymptotic justification to their extension.

2.3 A new approach involving empirical likelihood and resampling

In this section, we present a new approach, which involves empirical likelihood and resampling, to making inferences in the linear IV regression model. We first consider the parametric model in which $(u_i, \xi_i) \sim N(0, \Sigma)$. Under $H_0 : \beta = \beta_0$, there is no additional information derived from Π_2 . However, in a neighborhood of $\beta = \beta_0$, since one can learn about Π_2 from (17) up to $O_p(n^{-1/2})$ order, it can be used to supplement the least squares estimate of θ_2 based on (y_i, x_i, z_i) in the regression model (18). The likelihood ratio statistic assumes $\theta_2 = 0$ in the denominator and uses the MLE to estimate $(\theta_1, \theta_2, \Pi_1, \Pi_2, \Sigma)$ in the numerator. This is why we care about estimating $\theta_2 = \Pi_2'(\beta - \beta_0)$. On the other hand, the increase in information of the likelihood ratio statistic over the Anderson-Rubin statistic is negligible if the instrument is so weak that β cannot be consistently estimated, even though $\theta_2 = \Pi_2'(\beta - \beta_0)$ can be estimated with $O_p(\frac{1}{\sqrt{n}})$ error by regression under the standard conditions that $n^{-1} \sum_{i=1}^n (x_i', z_i')'(x_i', z_i')$ converges in probability to a nonrandom positive definite matrix and that (u_i, ξ_i') are i.i.d. with mean 0 and non-singular covariance matrix Σ . The “weak instrument” asymptotic theory developed during the past decade, beginning with seminal work of Staiger and Stock (1997) shows that under these standard regularity conditions, β cannot be consistently estimated if $\Pi_2 = O(n^{-1/2})$. In fact, in the case $\Pi_2 \sim n^{-1/2}C$ for some fixed matrix C , Staiger and Stock (1997) have shown that the IV estimator of β has a limiting nondegenerate distribution which depends on C and which differs from that of the LIML estimator. Since Π_2 is actually unknown, it has to be estimated from the regression model (17) of w_i on (x_i, z_i) . Under the preceding regularity conditions, the MLE or method of moments estimator $\hat{\Pi}_2$ differs from Π_2 by $O_p(n^{-1/2})$ and therefore $\hat{\Pi}_2 = O_p(n^{-1/2})$ cannot give a consistent estimate of C when $\Pi_2 \sim n^{-1/2}C$. This explains why (17) adds asymptotically negligible information to the regression model (18) when $\Pi_2 = O(n^{-1/2})$. On the other hand, if Π_2 is of a larger order of magnitude than $n^{-1/2}$, then $\hat{\Pi}_2 = (1 + o_p(1))\Pi_2$ and β can still be consistently estimated in the just- or over-identified case; see Lai, Small and Liu (2007a).

Since the asymptotic null distribution of the LR statistic depends on the unknown parameter matrix Π_2 when $\Pi_2 = O(n^{-1/2})$ and since there is little gain in information in using the LR statistic over the Anderson-Rubin statistic, a simple approach is to use a convex combination of both statistics. Note that the Anderson-Rubin statistic is a scalar multiple of the likelihood ratio statistic l_{LI} in the limited information (LI) model (18) whereas LR is the likelihood ratio statistic in the full information model (15)-(17). This suggests putting weight $h_n(\lambda_{\hat{\Pi}_2})$ to LR and $1 - h_n(\lambda_{\hat{\Pi}_2})$ to l_{LI} , where λ_M denotes the minimum eigenvalue of MM' for an $m_1 \times m_2$ matrix M , with $m_1 \leq m_2$, and h_n is a nondecreasing, continuously differentiable function with bounded derivative such that $h_n(x) = 0$ for $x \leq n^{-1}A_n$ and $h_n(x) = 1$ for $x \geq Kn^{-1}A_n$, with $K > 1$ and $A_n \rightarrow \infty$ slowly with n (e.g., $A_n \sim \log \log n$). Specifically, we use model averaging for the limited information and full information models to combine l_{LI} and LR into a modified likelihood ratio statistic of the form

$$l_{\text{mod}} = h_n(\lambda_{\hat{\Pi}_2})\text{LR} + (1 - h_n(\lambda_{\hat{\Pi}_2}))l_{\text{LI}}. \quad (19)$$

Although the distribution of l_{mod} depends on $h_n(\lambda_{\hat{\Pi}_2})$ and is therefore non-pivotal, $h_n(\lambda_{\hat{\Pi}_2}) = h_n(\lambda_{\Pi_2}) + O_p(n^{-1/2})$, the null distribution of l_{mod} can be consistently estimated because of the choice of $A_n \rightarrow \infty$ and the bounded derivative of h_n , under the assumption (made by Moreira and others) that $n^{-1} \sum_{i=1}^n z_i z_i'$ converges in probability to a positive definite matrix. Details are given in Lai, Small and Liu (2007), who use bootstrap resampling to estimate the $(1 - \alpha)$ -quantile of the distribution of l under $H_0 : \beta = \beta_0$. This bootstrap test of H_0 has an asymptotically correct type I error and has power comparable to that of the CLR test.

As pointed out in Section 2.2, the CLR test conditions on a component T of the bivariate invariant statistic so that the conditional null distribution of the likelihood ratio statistic given T is pivotal. Assuming the covariance matrix of (u'_i, ξ'_i) to be known, Moreira (2003, p. 1030) also notes that, “when β is known to equal β_0 , T is sufficient for Π_2 and is a one-to-one function of the constrained maximum likelihood estimator” of Π_2 that measures the strength of the instruments. Our approach uses the unconstrained MLE of Π_2 to form the weights in averaging between the full information and limited information models so that the null distribution of the modified likelihood ratio statistic (19), which is non-pivotal, can be consistently estimated. More importantly, this approach can be readily extended to the semi-parametric case in which (u'_i, ξ'_i) has an unspecified zero-mean distribution rather than a multivariate normal distribution with either specified or unknown covariance matrix Σ . The moment restrictions for this setting are $E[z_i(y_i - \beta'w_i - \gamma'x_i)] = 0$, $E[x'_i(y_i - \beta'w_i - \gamma'x_i)] = 0$, $E[y_i - \beta'w_i - \gamma'x_i] = 0$, $E[x'_i(w_i - \Pi_1 x_i - \Pi_2 z_i)] = 0$ and $E[z'_i(w_i - \Pi_1 x_i - \Pi_2 z_i)] = 0$. The full information empirical likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is the ratio of the maximum empirical likelihood over all $\beta, \gamma, \Pi_1, \Pi_2$ to the maximum empirical likelihood when β is set to β_0 . The limited information empirical likelihood ratio statistic is based on the regression model (18) instead. These two empirical likelihood ratio statistics can again be combined with (19), and, as in the parametric case, bootstrap resampling can be used to carry out

the test. Details are given in Lai, Small and Liu (2007).

2.4 Confidence Intervals

A commonly used method for constructing confidence intervals for univariate parameters β from IV or GMM estimators $\hat{\beta}$ is to use their asymptotic normality that yields approximate $(1 - \alpha)$ -level confidence intervals of the form $\hat{\beta} \pm \hat{\text{se}}\Phi^{-1}(1 - \frac{\alpha}{2})$, where $\hat{\text{se}}$ is an estimate of the standard error of $\hat{\beta}$ and Φ denotes the standard normal distribution function. For multivariate β , $\hat{\text{se}}$ is replaced by the estimated asymptotic covariance matrix of $\hat{\beta}$ to form confidence ellipsoids. Nagar (1959), Sargan (1975), Anderson and Sawa (1973, 1979) and Rothenberg (1983) have developed higher-order asymptotic expansions to improve the first-order normal approximation. However, in the case of weak instruments or moment restrictions, $\hat{\beta}$ may not even be consistent as noted above. Zivot, Startz and Nelson (1998) point out that this traditional approach produces confidence regions “that are highly misleading” when instruments are weak and give examples in which “traditional confidence regions *always* exclude the true parameters” in the presence of weak instruments and strong endogeneity. They propose alternative confidence regions based on inverting the Anderson-Rubin tests or other valid tests of $\beta = \beta_0$ so that β_0 is included in the $(1 - \alpha)$ -level confidence region if the level- α test accepts the null hypothesis $\beta = \beta_0$. Moreira (2003, Section 5) proposes to invert conditional likelihood ratio tests for the construction of confidence regions for β . We can likewise invert the modified empirical likelihood ratio test in Section 2.3 to form valid confidence regions without distributional assumptions on (u_i, ξ_i) ; see Lai, Small and Liu (2007a) for implementation details and simulation studies. Since the p -value of the test is computed by resampling, the confidence region thus constructed belongs to the class of *hybrid resampling* confidence regions introduced by Chuang and Lai (2000).

As pointed out by Chuang and Lai (2000), there are important differences between bootstrap and hybrid resampling confidence sets. Hybrid resampling in fact “hybridizes” (i) the exact method which assumes a completely specified model except for the unknown parameter of interest and which constructs the confidence set by inverting an exact test and (ii) the bootstrap method that replaces the unknown parameters by their sample estimates and the unknown distribution by the empirical distribution from which bootstrap resamples are drawn. For the moment restriction model (9), Hall and Horowitz (1996) have constructed bootstrap confidence intervals for β_0 by modifying the moment restrictions to accommodate the empirical distribution. Specifically, they note that the moment restrictions which are supposed to hold under F_0 , $E_{F_0}[g(X, \beta_0)] = 0$, do not usually hold with \hat{F}_n in place of F_0 for any value of β if there are more moment restrictions than parameters. To modify the usual bootstrap procedure in view of this, let $X_b^* = (X_1^*, \dots, X_b^*)$ be an i.i.d. sample from \hat{F}_n . Define $g^*(X, \beta) = g(X, \beta) - E_{\hat{F}_n}[g(X^*, \hat{\beta})]$, where $\hat{\beta}$ is the two-step GMM estimate based on (X_1, \dots, X_n) . Hall and Horowitz’s bootstrap version of the moment restriction $E_{F_0}[g(X, \beta_0)] = 0$ is $E_{\hat{F}_n}[g^*(X^*, \hat{\beta})] = 0$; the bootstrap version of the moment restriction is re-centered relative to the population version. Without this re-centering, the bootstrap would implement an incorrect moment restriction because generally $E_{\hat{F}_n}[g(X^*, \hat{\beta})] \neq 0$. Although they show that their bootstrap con-

fidence interval has two-sided coverage errors of order $o(n^{-1})$, Hall and Horowitz (1996) also note that the bootstrap does not provide more accurate finite-sample coverage in the presence of weak instruments. They define weak instruments as a setting in which $D \approx 0$, where $D = \frac{\partial}{\partial \beta} E_{F_0}[g(X, \beta_0)]$. If $D \approx 0$, the higher-order terms of the Edgeworth expansion of the studentized GMM statistic may dominate the lower-order terms for a given sample size and the bootstrap may provide little increase in accuracy or even make the errors worse (Hall and Horowitz, 1996). In Lai, Small and Liu (2007a), we present simulation evidence that shows the poor performance of the Hall-Horowitz bootstrap confidence intervals when there are weak instruments and show how these difficulties can be addressed by using hybrid resampling in lieu of the bootstrap.

3. Instrument and Model Selection

As noted in Section 2.3, the modified likelihood ratio statistic (19) is obtained by averaging two models that relate the observed data to β , namely, the full information model (15)-(17) and the limited information model (18). Because of the nature of h_n , model averaging actually amounts to model selection in the case $h_n(\lambda_{\hat{\Pi}_2}) \leq A_n/n$ (for which the limited information model is chosen) and in the case $h_n(\lambda_{\hat{\Pi}_2}) \geq K A_n/n$ (for which the full information model is chosen). Note that both models are assumed correct in Section 2.3 and that model averaging or model selection is targeted towards estimation of β . In this section, we consider the more general problem of joint model and instrument (moment restriction) selection from a set of plausible models and moment restrictions, some of which may not be valid. We first review previous approaches and then describe a new approach that is targeted towards the estimation of β . This approach is similar in philosophy to the *focused information criterion* proposed by Claeskens and Hjort (2003) and it consists of two stages. The first stage eliminates models that are too far from the true model as measured by the empirical likelihood ratio. The second stage chooses, among those models not yet eliminated, a model that gives the best precision for estimating the parameter of interest in the econometric application.

3.1 Pretesting Approach

The traditional econometric approach to addressing the problems of model and moment selection in IV regression or moment restriction models is “pretesting” combined with a sequential search strategy. This approach is analogous to traditional stepwise variable selection in regression. The researcher starts with a model and a set of moment restrictions, uses some procedure to propose a change to the model or the moment restrictions and then decides whether to make this change based on a hypothesis test. The null hypothesis is typically that the more restrictive model is correct or that the larger set of moment restrictions is valid and the procedure is typically stopped once a null hypothesis is rejected. See Hall (2005) for a review of methods for testing that a subset of moment restrictions is valid under the assumption that the remaining moment restrictions in the full set are valid. One particular test is Eichenbaum, Hansen and Singleton’s (1988) GMM test which uses the test statistic

$$T = \min_{\beta} \hat{G}'_{H_0}(\beta) [\hat{\text{Cov}}_{H_0}(g_{H_0}(X, \tilde{\beta}_{H_0}))]^{-1} \hat{G}_{H_0}(\beta) - \min_{\beta} \hat{G}'_{H_1}(\beta) [\hat{\text{Cov}}_{H_1}(g_{H_1}(X, \tilde{\beta}_{H_1}))]^{-1} \hat{G}_{H_1}(\beta),$$

where g_{H_0} represents the full set of moment restrictions that are valid under the null hypothesis H_0 and g_{H_1} represents the restricted set of moment restrictions that are valid under the alternative hypothesis H_1 . As in Section 2.1, for $j = 0, 1$, $\hat{G}_{H_j}(\beta)$ denotes $\frac{1}{n} \sum_{i=1}^n g_{H_j}(X_i, \beta)$ and $\tilde{\beta}_{H_j}$ is an initial consistent GMM estimator under H_j that uses the identity weight matrix. Under H_0 that the full set of moment restrictions is valid, the test statistic T converges in distribution to a χ^2 distribution with degrees of freedom equal to the number of extra moment restrictions in H_0 compared to H_1 .

Sequential search strategies are often applied in an informal manner by empirical researchers (Andrews and Lu, 2001). As a formalization of this informal procedure, Andrews and Lu discuss the following downward testing procedure. Consider a set of model and moment selection vectors $\{(b_j, c_j), j = 1, \dots, M\}$, where $|b_j|$ denotes the number of parameters to be estimated by the model and $|c_j|$ denotes the number of moment restrictions used to estimate these parameters; see Section 3.2 for a more detailed description. Starting with the model and model selection vectors for which $|c| - |b|$ is the largest, carry out GMM tests of overidentifying restrictions with progressively smaller $|c| - |b|$ until a test is found which does not reject the null hypothesis that the moments c_j considered are all valid for the given model b_j .

In the statistical literature on model selection for regression, sequential search strategies based on hypothesis tests have been criticized on several grounds. First, the selection of significance levels is necessarily subjective and their interpretation is unclear. It seems preferable to use a more decision-theoretic model selection approach that addresses the problem of choosing the best model for the goals of the analysis directly. Second, sequential hypothesis testing approaches use a local rather than a global search in looking for the best model and may choose a model that is far from optimum under any reasonable model selection criterion. A related drawback is that the sequential hypothesis testing approaches choose one model rather than providing an overall picture of the most promising models. See Linhart and Zucchini (1986), Miller (1990) and Bickel and Zhang (1992) for further discussion of these points.

3.2 Criterion-based approach

A criterion-based approach to model selection addresses the weaknesses of pretesting. A criterion is chosen to reflect how well the model fits the data for the envisaged purposes of the data analysis. The best model or a list of the few most promising models can then be chosen according to this criterion. One reasonable way to form a criterion is to first choose a discrepancy that measures the lack of fit between the fitted model and the true model for the purposes of the data analysis (Linhart and Zucchini, 1986). The criterion can then be an estimate of the discrepancy. Mallows' C_p statistic and Akaike's Information Criterion (AIC) are examples of this approach (Akaike, 1969; Mallows, 1973).

Andrews and Lu (2001) develop a criterion-based approach to model and moment selection for GMM estimation in moment restriction models. Similar in spirit to Andrews and Lu's approach, Hong, Preston and Shum (2003) develop a criterion-based approach that uses generalized empirical likelihood statistics rather than GMM statistics. These papers are important advances in model selection for semi-parametric moment restriction models (Hansen, 2005). We focus on Andrews and Lu's criteria to explain the basic ideas.

Andrews and Lu's criteria are based on an assessment of whether the given moment restrictions are valid for the given model. Let (b, c) denote a pair of model and moment selection vectors whose components are 1 or 0. If the j th element of b is 1, then the j th element of the parameter vector θ is to be estimated. If the j th element of b is 0, then the j th element of θ is not estimated and set equal to some prescribed value, which can be denoted by 0 (after reparameterization if necessary). If the j th element of c is 1, then the j th moment restriction is included in the GMM criterion function, whereas if the j th element is 0, it is not included. Let $\Theta_{[b]}$ denote the set of parameter vectors $\theta_{[b]}$ whose components are set equal to 0 whenever b sets them to 0. Let $\hat{G}_c(\theta)$ denote the vector of sample moment conditions selected by c . The GMM estimator based on the model selected by b and the moment restrictions selected by c is

$$\hat{\theta}_{b,c} = \arg \min_{\theta_{[b]} \in \Theta_{[b]}} \hat{G}'_c(\theta_{[b]}) W_n(b, c) \hat{G}_c(\theta_{[b]}), \quad (20)$$

where $W_n(b, c)$ is the weight matrix employed with the moments $\hat{G}_c(\theta_{[b]})$ and the model selected by b , e.g., $W_n(b, c)$ is the inverse of the sample covariance of the moment restrictions based on an initial estimator $\tilde{\theta}_{b,c}$ that minimizes (20) with the identity matrix replacing $W_n(b, c)$. Denote by $J_n(b, c)$ the value of the GMM objective function at $\hat{\theta}_{b,c}$, i.e.,

$$J_n(b, c) = \hat{G}'_c(\hat{\theta}_{b,c}) W_n(b, c) \hat{G}_c(\hat{\theta}_{b,c}).$$

Let $|b|$ and $|c|$ denote the number of parameters of θ selected by b and the number of moments selected by c , respectively. Also let \mathcal{BC} be the space of parameter and moment selection vectors considered. Andrews and Lu propose to choose the selection pair $(b, c) \in \mathcal{BC}$ that minimizes the criterion

$$J_n(b, c) + (|b| - |c|)f(n), \quad (21)$$

for which they propose several choices of $f(n)$: (i) $f(n) = 2$; (ii) $f(n) = Q \ln \ln n$, $Q > 2$; and (iii) $f(n) = \log n$. The first term $J_n(b, c)$ can be considered as a test statistic for testing whether $E[\hat{G}_c(\theta_{[b]})] = 0$ for some $\theta_{[b]} \in \Theta_{[b]}$. The test statistic $J_n(b, c)$ has a $\chi^2_{|c|-|b|}$ distribution under the null hypothesis that the moment restrictions are valid; see Hall (2005). The second term in (21) is the negative of $(|c| - |b|)f(n)$, which penalizes model selection vectors that use more parameters and rewards moment selection vectors that use more moment conditions. Note that if the moment

restrictions of (b, c) and (b', c') are both valid and $|c| - |b| > |c'| - |b'|$, then (b, c) will be preferred because (for $n > 5$)

$$\begin{aligned} & E[J_n(b, c) + (|b| - |c|)f(n) - J_n(b', c') - (|b'| - |c'|)f(n)] \\ &= [(|b| - |c|) - (|b'| - |c'|)](f(n) - 1) < 0. \end{aligned}$$

According to the first-order asymptotic theory of GMM, it is best (or at least it cannot hurt) to use all valid moment restrictions. Andrews and Lu's criterion is sensible according to this first-order theory because it rewards using more moment restrictions. When $f(n) \rightarrow \infty$ and $f(n) = o(n)$, their criterion is consistent in the following sense. Let \mathcal{L} denote the set of selection pairs $(b, c) \in \mathcal{BC}$ such that the moment restrictions are satisfied for model b , i.e., there exists $\theta_{[b]} \in \Theta_{[b]}$ such that $E[\hat{G}_c(\theta_{[b]})] = 0$. Suppose that there exists a parameter/moment selection pair $(b^*, c^*) \in \mathcal{L}$ such that

$$|c^*| - |b^*| > |c| - |b| \text{ for every } (b, c) \in \mathcal{L}, (b, c) \neq (b^*, c^*) \quad (22)$$

Then the probability that Andrews and Lu's criterion selects the model and moment pair (b^*, c^*) converges to 1.

3.3 Econometric issues and an example

To illustrate the methods in Sections 3.1 and 3.2 and certain related issues, we consider the dynamic panel data model (2) without covariates, in which the parameter β is of primary economic interest. For example, in Lee et al. (1998), β determines the rate of convergence of the economic outputs of countries. Using differencing as in Anderson and Hsiao (1981), Arellano and Bond (1991) consider moment restrictions of the form

$$E(u_{it}) = 0 \text{ for } 1 \leq t \leq T, \quad (23)$$

$$E(y_{is}\Delta u_{it}) = 0 \text{ for } 2 \leq t \leq T, \quad 0 \leq s \leq t - 2. \quad (24)$$

If the individual time series $\{y_{i0}, \dots, y_{iT}\}$ are stationary, then we can include an additional moment restriction

$$E[(u_{i2} + \dots + u_{iT})\Delta y_{i1}] = 0. \quad (25)$$

We now present a simulation study of the pretesting and Andrews-Lu approaches to choosing between model (23)-(24) and model (23)-(25) in which we consider the following two designs:

- (I) The individual time series are stationary and $\beta = 0.9$ so that each individual time series is highly persistent. The initial conditions are

$$y_{i0} = \frac{\alpha_i}{1 - \beta} + v_{i0},$$

Table 1

Comparison of performance of four moment selection methods in Design I

Estimator	$\sqrt{\text{MSE}}$			Proportion Correct Model Selected		
	n=100	n=250	n=500	n=100	n=250	n=500
Andrews and Lu	0.519	0.396	0.271	0.897	0.943	0.974
Pretesting	0.530	0.422	0.282	0.877	0.904	0.936
Always Use (24)	0.682	0.559	0.398	0	0	0
Always Use (24) and (25)	0.446	0.358	0.258	1	1	1

Table 2

Comparison of performance of four moment selection methods in Design II

Estimator	$\sqrt{\text{MSE}}$			Proportion Correct Model Selected		
	n=100	n=250	n=500	n=100	n=250	n=500
Andrews and Lu	0.160	0.095	0.061	0.938	0.977	0.975
Pretesting	0.160	0.095	0.061	0.919	0.944	0.934
Always Use (24)	0.163	0.095	0.058	0	0	0
Always Use (24) and (25)	0.159	0.095	0.061	1	1	1

where $v_{i0} \sim N(0, \frac{\sigma_u^2}{1-\beta^2})$.

(II) The individual time series are stationary and $\beta = 0.5$.

For each design, $T = 5$ and $(\alpha_i, u_{i,-1}, \dots, u_{iT}) \sim N(0, I)$. We consider four estimators of β : (i) the estimator of β obtained by applying Andrews and Lu procedure with $f(n) = \log n$, (ii) pretesting to decide on the model where we test the null hypothesis that the additional moment restriction (25) is true at a nominal 0.05 level using the GMM test statistic of Eichenbaum, Hansen and Singleton (1988), (iii) always use only the moment restriction (23)-(24) and (iv) always uses the full set of moment restrictions (23)-(25). For both designs, the “correct” model has the full set of moment restrictions (23)-(25). Tables 1 and 2 show the root mean squared errors of the four estimators and the proportion of times the associated model selection method chose the correct model. Each result is based on 1000 simulations.

Table 1 shows that in Design I, the estimator that uses restrictions (23)-(25) does substantially better than the estimator that uses only (23)-(24), but Table II shows that the differences are negligible in Design II. In statistical model selection studies, the right portion of Tables 1 and 2 which displays the proportion of times the correct model was selected is an important measure of the performance of a model selection criterion. But for the econometric problem we are studying, model selection is only a device for finding the best estimator of β , the parameter of primary interest, and the proportion of times the correct model was selected may not be relevant to this goal.

A challenging issue that arises in many models in econometrics is related to

weakly informative moment restrictions, such as the weak instruments discussed in Section 2. A weakly informative moment restriction is a moment restriction that provides little information relative to the sample size about the parameter of interest. Quantitative measures of the weakness of moment restrictions are discussed by Stock and Wright (2000). In Design I, where $\beta = 0.9$, the moment restrictions (24) are weakly informative but the stationarity moment restriction (25) is highly informative about β . In contrast, in Design II, where $\beta = 0.5$, the moment restrictions (24) are more informative and the stationarity moment restriction (25) is less informative. There is substantial evidence that the use of weakly informative moment restrictions or too many moment restrictions causes the performance of GMM to deteriorate (e.g., Altonji and Segal, 1996; Andersen and Sorensen, 1996; Podivinsky, 1999; Stock and Wright, 2000). As an example, for Design I in our simulation study where the moment restrictions (24) are weakly informative, the estimator A which uses only the moment restrictions (23) and (25) and ignores the valid but weakly informative moment restrictions (24) performs substantially better than the estimator B that uses the full set of valid moment restrictions (23)-(25): the root mean squared errors for the estimator A are 0.144, 0.140 and 0.157 for $n = 100, 250$ and 500 respectively, compared to 0.446, 0.358 and 0.258 for the estimator B . On the other hand, for Design II, the estimator B is substantially better with root mean squared errors of 0.159, 0.095 and 0.061 for $n = 100, 250$ and 500 respectively, compared to 0.378, 0.345 and 0.255 for the estimator A . Andrews and Lu's model selection criteria are aimed at including all moment restrictions that appear valid, and are not well suited for settings with weak instruments in which we might want to only make use of moment restrictions that are not only valid but also not weak. In particular, we would like be able to choose the estimator A when it performs better than B . Lai, Liu and Small (2007) recently developed a model selection method that addresses this issue.

3.4 *A two-stage approach involving empirical likelihood and resampling*

Lai, Small and Liu's (2007b) approach is motivated by the following two considerations. First, we would like to eliminate from consideration those models that do not approximate well the true model. Second, among those models and moment restrictions which are not eliminated, we would like to choose the model and moment restriction combination which provides the best estimate of the parameter vector of interest. These considerations lead to the following two-stage procedure.

Suppose that we observe i.i.d. d -dimensional random vectors z_i , $1 \leq i \leq n$, with common distribution F_0 , and that we are given a family \mathcal{P} of moment restriction models that involve a p -dimensional parameter vector θ . Of primary econometric interest is estimation of an s -dimensional subvector θ_* of θ_0 . Therefore, which moment restriction model in \mathcal{P} yields the best model-based estimator of θ_* is the associated model selection problem. Let $\mathcal{Q}_m = \bigcup_{\theta \in \Theta_m} \mathcal{Q}_m(\theta)$ be subsets of \mathcal{P} , where the members of \mathcal{Q}_m have an r_m -dimensional vector of moment restrictions $E[g(z, \theta)] = 0$, with $g : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^{r_m}$, and Θ_m is a $(p - q_m)$ -dimensional subset of Θ that sets certain components of θ to 0, as in Section 3.2.

Our two-stage approach to model and moment selection proceeds as follows. In

the first stage, for model \mathcal{Q}_m with associated maximum empirical likelihood estimator $\hat{\theta}_m$, compute the empirical log-likelihood ratio $l_m = 2\{n \log n - \log L_{\mathcal{Q}_m}(\hat{\theta}_m)\}$, where $L_{\mathcal{Q}_m}(\theta)$ is the empirical likelihood of θ under \mathcal{Q}_m . We eliminate \mathcal{Q}_m from consideration if

$$l_m > \max\{(r_m - p + q_m)n\epsilon_n, \chi_{r_m-p+q_m;1-\alpha}^2\} \quad (26)$$

where $n\epsilon_n \rightarrow \infty$ and $\epsilon_n \rightarrow 0$. This ensures that wrong models will consistently be eliminated and correct models will consistently not be eliminated. A typical choice of ϵ_n is $\frac{\log n}{2n}$, in analogy with BIC. As explained in Lai, Liu and Small (2007), the rationale behind (26) is related to both the asymptotic χ^2 -distribution and the theory of moderate deviations for empirical likelihood ratio statistics.

The second stage consists of choosing, among the models which have not been eliminated from consideration in the first stage, the associated estimator $\hat{\theta}_m$ which minimizes an estimate of a measure of the size of $\text{Cov}(\hat{\theta})$. The size of Cov is measured in a way that is suited to the goals of the estimation problem. If we are interested in one particular component θ_* of θ , then we use the variance of the estimator of that component. If we are interested in a subvector θ_* , then we use $\text{trace}\{\text{Cov}(\hat{\theta}_{*,m})\}$ or $\det\{\text{Cov}(\hat{\theta}_{*,m})\}$. We use bootstrap methods to estimate $\text{Cov}(\hat{\theta}_m)$. The bootstrap estimate of the estimator's "dispersion size" is able to better reflect its finite sample properties than the first-order asymptotic theory when there are weakly informative moment restrictions or many moment restrictions. Further details on the method and a study of its performance are provided in Lai, Liu and Small (2007).

References

- Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. of Statist. Math.* **21**: 243-247.
- Alonso-Borrego, C., Arellano, M. (1999). Symmetrically normalized instrumental-variable estimation using panel data. *J. Bus. Economic Statist.* **17**(1): 36-49.
- Altonji, J., Segal, L. (1996). Small sample bias in GMM estimation of covariance structures. *J. Bus. Economic Statist.* **14**: 353-366.
- Andersen, T.G., Sørensen, B.E. (1996). GMM estimation of a stochastic volatility model: a Monte Carlo study. *J. Bus. Economic Statist.* **14**: 328-352.
- Anderson, T.W. (1950). Estimation of parameters of a single equation by the limited-information maximum likelihood method. In *Statistical Inference in Dynamic Economic Models*, T.C. Koopmans, ed., Cowles Commission for Research in Economics, Wiley, New York: 311-322.
- Anderson, T.W., Hsiao, C. (1981). Estimation of dynamic models with error components. *J. Amer. Statist. Assoc.* **76**: 598-606.
- Anderson, T.W., Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *J. Econometrics* **18**: 47-82.
- Anderson, T.W., Rubin, H. (1949). Estimation of the parameters of coefficients of a single equation in a simultaneous system and their asymptotic expansions. *Econometrica* **41**: 683-714.
- Anderson, T.W., Rubin, H. (1950). The asymptotic properties of estimates of the

- parameters of a single equation in a complete system of stochastic equations. *Ann. Mathem. Statist.* **21**: 570-582.
- Anderson, T.W., Sawa, T. (1973). Distribution of estimations of coefficients of a single equation in a simultaneous system and their asymptotic expansion. *Econometrica* **41**: 683-714.
- Anderson, T.W., Sawa, T. (1979). Evaluation of the distribution function of the two-stage least squares estimate. *Econometrica* **47**: 163-182.
- Andrews, D.W.K., Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econometrics* **101**: 123-164.
- Andrews, D.W.K., Moreira, M., Stock, J.H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* **74**: 715-752.
- Andrews, D.W.K., Moreira, M., Stock, J.H. (2007). Performance of conditional Wald tests in IV regression with weak instruments. *J. Econometrics* **139**: 116-132.
- Angrist, J.D., Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? *Quar. J. Economics* **106**: 979-1014.
- Arellano, M., Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Economic Studies* **58**: 277-297.
- Balestra, P., Nerlove, M. (1966). Pooling cross-section and time series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica* **34**: 585-612.
- Baltagi, B.H., Levin, D. (1986). Estimating dynamic demand for cigarettes using panel data: the effects of bootlegging, taxation and advertising reconsidered. *Rev. Economics Statist.* **48**: 148-155.
- Bhargava, A. (1991a). Estimating short and long run income elasticities of foods and nutrients for rural South India. *J. Royal Statist. Soc. A* **154**: 157-174.
- Bhargava, A. (1991b). Identification and panel data models with endogenous regressors. *Rev. Economic Studies* **58**: 129-140.
- Bickel, P.J., Zhang, P. (1992). Variable selection in nonparametric regression with categorical covariates. *J. Amer. Statist. Assoc.* **87**: 90-97.
- Blundell, R., Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *J. Econometrics* **87**: 115-143.
- Bound, J., Jaeger, D.A., Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90**: 443-450.
- Burnside, C. (1996). Production function regressions, returns to scale and externalities. *J. Monetary Economics* **37**: 177-201.
- Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Mathem. Statist.* **23**(4): 493-507.
- Chiang, C.L. (1956). On regular best asymptotically normal estimates. *Ann. Mathem. Statist.* **27**: 336-351.
- Chuang, C.-S., Lai, T.L. (2000). Hybrid resampling methods for confidence intervals. *Statist. Sinica* **10**: 1-33.

- Claeskens, G., Hjort, N.L. (2003). The focused information criterion (with discussion). *J. Amer. Statist. Assoc.* **98**: 879-899.
- Donald, S.G., Newey, W.K. (2001). Choosing the number of instruments. *Econometrica* **69**(5): 1161-1191.
- Eichenbaum, M., Hansen, L.P., Singleton, K.S. (1988). A time series analysis of representative agent models of consumption and leisure under uncertainty. *Quar. J. Economics* **103**: 51-78.
- Ferguson, T.S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Statist.* **29**: 1046-1062.
- Gallant, A.R., Tauchen, G. (1996). Which moments to match? *Econometric Theory* **12**: 657-681.
- Hall, A.R. (2005). *Generalized Method of Moments*. Oxford University Press, New York.
- Hall, P., Horowitz, J. (1996). Bootstrap critical values for tests based on generalized method of moments. *Econometrica* **64**: 891-916.
- Hansen, B. (2005). Challenges for econometric model selection. *Econometric Theory* **21**: 60-68.
- Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4): 1029-1054.
- Harris, D., Matyas, L. (1999). Introduction to the generalized method of moments estimation. In *Generalized Method of Moments Estimation* (L. Matyas, ed.). Cambridge University Press, New York.
- Hong, H., Preston, B., Shum, M. (2003). Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory* **19**: 923-943.
- Hsiao, C. (1985). Benefits and limitations of panel data. *Econometric Reviews* **4**(1): 121-174.
- Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* **69**: 1661-1672.
- Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* **70**: 1781-1804.
- Kolaczyk, E.D. (1995). An information criterion for empirical likelihood with general estimating equations. Technical Report 417, University of Chicago Department of Statistics.
- Lai, T., Liu, J. and Small, D. (2007). Efficient Model and Moment Selection for Moment Restriction Models with Application to Dynamic Panel Data. Tech. report, Department of Statistics, Stanford University.
- Lai, T., Small, D. and Liu, J. (2007). Valid inference with weak instruments via empirical likelihood and resampling. Tech. report, Department of Statistics, Stanford University.
- Lee, M., Longmire, R., Mátyás, L., Harris, M. (1998). Growth convergence: some panel data evidence. *Applied Economics* **30**: 907-912.
- Lee, C.-I. (2001). Finite sample bias in IV estimation of intertemporal labor supply models: is the intertemporal substitution elasticity really small? *Rev. Economic*

- Stat.* **83(4)**: 638-646.
- Linhart, H., Zucchini, W. (1986). *Model Selection*. John Wiley & Sons, New York.
- Mallows, C. (1973). Some comments on C_p . *Technometrics* **15**: 661-675.
- Miller, A. (1990). *Subset Selection in Regression*. Chapman & Hall, London.
- Moreira, M. (2003). A conditional likelihood ratio test for structural models. *Econometrica* **71**: 1027-1048.
- Nagar, A.L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* **27**: 575-595.
- Nelson, C.R., Startz, R. (1990a). The distribution of the instrumental variable estimator and its t -ratio when the instrument is a poor one. *J. Business* **63**: 125-140.
- Nelson, C.R., Startz, R. (1990b). Some further results on the exact small properties of the instrumental variable estimator. *Econometrica* **58(4)**: 809-837.
- Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**: 237-249.
- Pesaran, M.H., Haque, N.U., Sharma, S. (2000). Neglected heterogeneity and dynamics in cross-country savings regressions. In *Panel Data Econometrics - Future Direction: Papers in Honor of Pietro Balestra*. Elsevier Science, Amsterdam.
- Pesaran, M.H., Smith, R.J. (1994). A generalized R^2 criterion for regression models estimated by the instrumental variable method. *Econometrica* **62**: 705-710.
- Podivinsky, J.M. (1999). Finite sample properties of GMM estimators and tests. In *Generalized Method of Moments Estimation* (L. Mátyás, ed.). Cambridge University Press, Cambridge.
- Rothenberg, T.J. (1983). Asymptotic properties of some estimators in structural models. In *Studies in Econometrics, Time Series, and Multivariate Statistics* (S. Karlin, T. Amemiya and L. Goodman, eds.) Academic Press.
- Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**: 300-325.
- Ramalho, J., Smith, R. (2002). Generalized empirical likelihood non-nested tests. *J. Econometrics* **102**: 1-28.
- Sargan, J.D. (1975). Asymptotic theory and large models. *Int. Economic Rev.* **16(1)**: 75-91.
- Smith, R. (1992). Non-nested tests for competing models estimated by generalized method of moments. *Econometrica* **60**: 973-980.
- Staiger, D., Stock, J.H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65(3)**: 557-586.
- Stock, J.H., Wright, J.H. (2000). GMM with weak identification. *Econometrica* **68(5)**: 1055-1096.
- Taamouti, M. (2001). Techniques of exact inference in structural models with macroeconomic applications. Ph.D. dissertation. University of Montreal, Department of Economics.
- Wang, J., Zivot, E. (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* **66**: 1389-1404.
- Zivot, E., Startz, R., Nelson, C.R. (1998). Valid confidence intervals and inference in the presence of weak instruments. *Int. Economic Rev.* **39**: 1119-1144.